Web appendix 2:

SUPPLEMENTARY MATERIALS

Efficacy and safety of betahistine treatment in patients with Menière's disease: primary results of a long term, multicentre, double blind, randomised, placebo controlled, dose defining trial (*BEMED* trial)

Christine Adrion, Carolin Simone Fischer, Judith Wagner, Robert Gürkov, Ulrich Mansmann, Michael Strupp [#]; for the *BEMED* study group

[#] Corresponding author: E-mail Michael.Strupp@med.uni-muenchen.de

The PDF file includes:

S1. Investigators and Participating Centres

17 centres (Neurology or ENT departments of university hospitals) screened for eligible patients.

14 of them allocated 221 study participants.

*The following are members of the **BEMED Study Group**:*

Prof. Dr. Michael Strupp, Dr. Carolin Simone Fischer; Department of Neurology, University Hospital Munich, Campus Grosshadern, Munich, Germany

Prof. Dr. Eike Krause, PD Dr. Robert Gürkov; Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Munich, Campus Grosshadern, Munich, Germany

Dr. Sabrina Holzapfel; Hals-Nasen-Ohrenklinik und Poliklinik, Klinikum rechts der Isar der Technischen Universität München, Munich, Germany

Prof. Dr. Martin Westhofen; Department of Otorhinolaryngology and Head and Neck Surgery, University Hospital Aachen, RWTH Aachen University, Aachen, Germany

Prof. Dr. Thomas Lempert; Department of Neurology, Schlosspark-Klinik, Berlin, Germany

Prof. Dr. Hubert Löwenheim; Department of Otorhinolaryngology - Head & Neck Surgery, Hearing Research Center, University of Tübingen Medical School, Tübingen, Germany

PD Dr. Michael v. Brevern; Department of Neurology, Park-Klinik Weißensee, Berlin, Germany

Prof. Dr. Thomas Lenarz; Department of Otolaryngology, Hannover Medical School, Hannover, Germany

Prof. Dr. Hans-Christoph Diener; Department of Neurology and Stroke Center, University Hospital Essen, Essen, Germany

Dr. Hermann Hilber; Department of Otorhinolaryngology, University Hospital Regensburg, Regensburg, Germany

Dr. Ines Repik; Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital Mannheim, Medical Faculty Mannheim of the Heidelberg University, Mannheim, Germany

Dr. Türker Basel, Dr. Daniel Weiß; Department of Otorhinolaryngology, Head and Neck Surgery, University Hospital of Muenster, Muenster, Germany

Prof. Dr. Dirk Beutner; Department of Otorhinolaryngology, Head and Neck Surgery, University of Cologne, Medical Faculty, Cologne, Germany

PD Dr. Holger A. Rambold; Department of Neurology, County Hospitals of Altötting and Burghausen, Altötting, Germany

S2. Procedural and Statistical Methods

2.1 Sensitivity analyses

2.1.1 GLM for time interval {7, 8, 9}

To deal with missing values a sensitivity analysis of the primary efficacy outcome was performed which only used patients with a total number of evaluated days larger than 0 within the 90-day assessment period (months 7, 8 and 9). This particular MAR-based analysis examined whether patients who withdrew before time interval 7 showed comparable efficacy results with respect to the overall primary analysis. Marked differences would indicate a strong selection process and informative missingness. This pre-planned sensitivity analysis excluded patients who withdrew totally from the study before time interval 7. The simple negative binomial model (NB GLM) was based on an aggregated version of the longitudinal approach used for the main model by summarizing the number of Menière's attacks and the number of evaluated days within time intervals 7, 8, and 9 only (should be 90 days according to the protocol). The linear predictor for the generalized linear model was defined according to the mixed effects model chosen for the primary analysis, leaving out the random effects part and time effects.

2.1.2 Supportive efficacy analyses (unplanned): Types of vertigo attacks

As supportive post-hoc primary efficacy analyses and in order to substantiate the robustness of the estimated treatment effect, we explored two alternative (stricter) definitions of a *Menière's attack* derived from the raw daily patient-reported vertigo symptoms. This restriction implied that rather mild types of patient-reported vertigo symptoms classified as gait unsteadiness and/or lightheadedness (hence without the criteria rotatory and/or postural documented on the original patient diary) were ignored, because they were assumed to have a potential diluting effect.

The incidence of episodes of vertigo classified as rotatory and/or postural ("*RP-attacks*"), and rotatory ("*R-attacks*") were analyzed in an exploratory fashion to investigate the diminishment over time and

its relation to intervention by using exactly the same definition concerning time units (30-day intervals) and the time at risk for attacks (number of evaluated days per 30-day interval). Therefore, the methodological concept applied for the primary efficacy analysis was adopted for these two derived efficacy endpoints in an analogous manner.

2.2 Missingness Map

The proportion of missings over time (i.e. per 30-day interval 1, 2, …, 9) can be visualized by means of a so-called missingness map: This plot displays the missingness patterns concerning diary information across time interval 1 to 9. In the case of missing data for at least one 30-day interval, corresponding gaps are indicated in light grey; time periods with available data are displayed in dark grey. The proportion of intermittent missings was rather low for each treatment arm; the proportion of monotone missing attack information was lower in the LD group. Altogether, the proportion of monotone (and intermittent) missings was not higher than expected for symptomatic trials assessing the ability of an intervention to provide symptom relief from the condition.

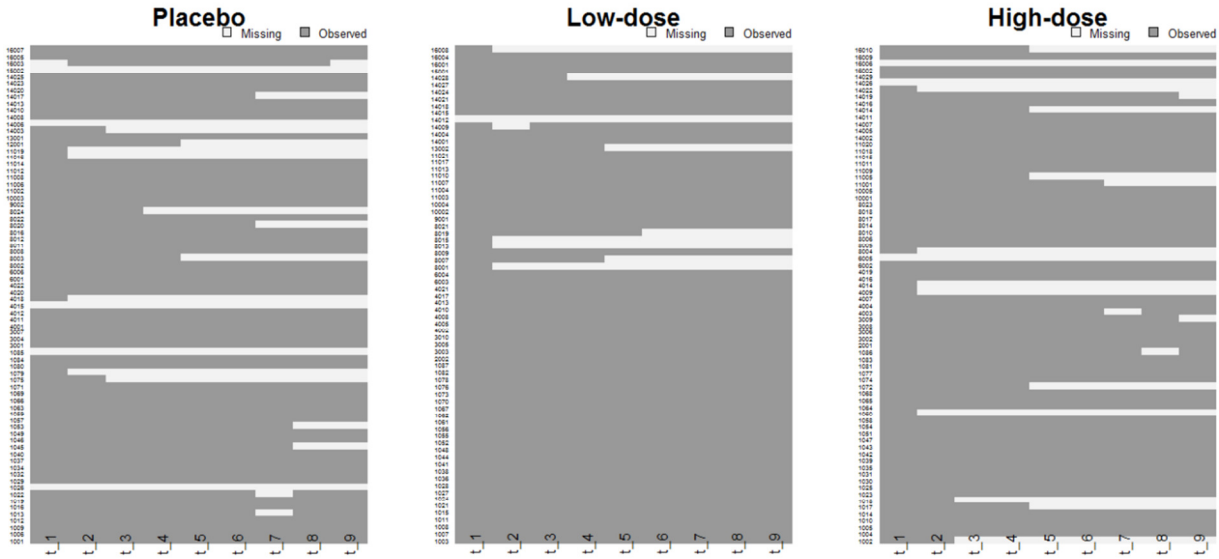

Figure S1. Missingness Maps showing where missingness occurs in the attack dataset (ITT sample) within the time period of primary interest (interval 7, 8, 9). The figures display monotone and intermittent missing data patterns concerning diary information of the ITT population. Dark grey: diary data available; light grey: missing data.
x-axis: number for each 30-day interval 1, 2, …, 9. y-axis: observation number representing 219 out of 221 PatIDs (patients are sorted according to their original PatID within each treatment group). 2 low-dose patients were excluded since they did not provide any post-randomization data.

2.3 Additional efficacy analyses: adjusting for centre effects, subgroup analyses

Centre effects. Study site was used as stratification variable in the allocation process. Since the BEMED trial was not explicitly designed with enough power to detect centre effects, the primary efficacy analysis was performed unadjusted. Centre was studied as one of the exploratory adjusted analyses by pooling of small sites with fewer than 15 randomized patients (based on geographical considerations), and, additionally, pooling of sites located in Munich within the catchment area of the German Center for Vertigo and Balance Disorders (DSGZ), which recruited the largest number of patients. For these preplanned adjusted analyses the linear predicator of the primary model was extended by including the additional fixed effects terms `site*time` and `site*time*trtgroup`. Afterwards, a formal statistical interaction test was applied.

Covariates gender and age. According to the main efficacy analysis, pre-specified subgroup effects were explored by including interaction terms between treatment group and the baseline covariates gender and age, the latter with pre-specified cut-off points ≤45, (45, 55], (55, 65], >65 years. These exploratory subgroup analyses focused on the evidence for a difference in treatment effects, investigating for potential interaction effects.

2.4 Multiplicity Issues

All null hypotheses were tested at the nominal two-sided 5% significance level.
HD, LD and PL groups were compared in terms of the primary endpoint by using a formal closed-testing procedure that examines the three hypotheses with respect to the three comparisons HD vs. LD, HD vs. PL, and LD vs. PL by preserving the overall 5% significance level of the confirmatory efficacy analyses. The closed-testing procedure (Marcus et al., 1976; Bauer 1991) consisted of overall global test testing of whether there is any treatment effect at all (referring to the omnibus treatment-by-time interaction), followed by three pairwise comparisons using the same significance level of 5%. If the global test for the global null hypothesis was not significant no pairwise comparisons would be valid. The likelihood ratio (LR) test was performed as a global test.

The secondary outcomes were analyzed in an exploratory manner, and the results are only interpreted as supportive evidence related to the primary efficacy outcome. It was of interest to investigate if the results also indicate a lack of a beneficial effect between the treatment groups.

2.5 Plot for marginal mean attack rates per 30 days at month 7 to 9

The following figure visualizes Table 4 of the main article.

The principal study question and analytic objective of the BEMED trial was to quantify the *(marginal) average of the mean attack rates* across the pre-specified 90-day assessment period (month 7, 8, 9) at the end of the entire 9-month treatment period, and to compare between the three treatment groups PL, LD, and HD.

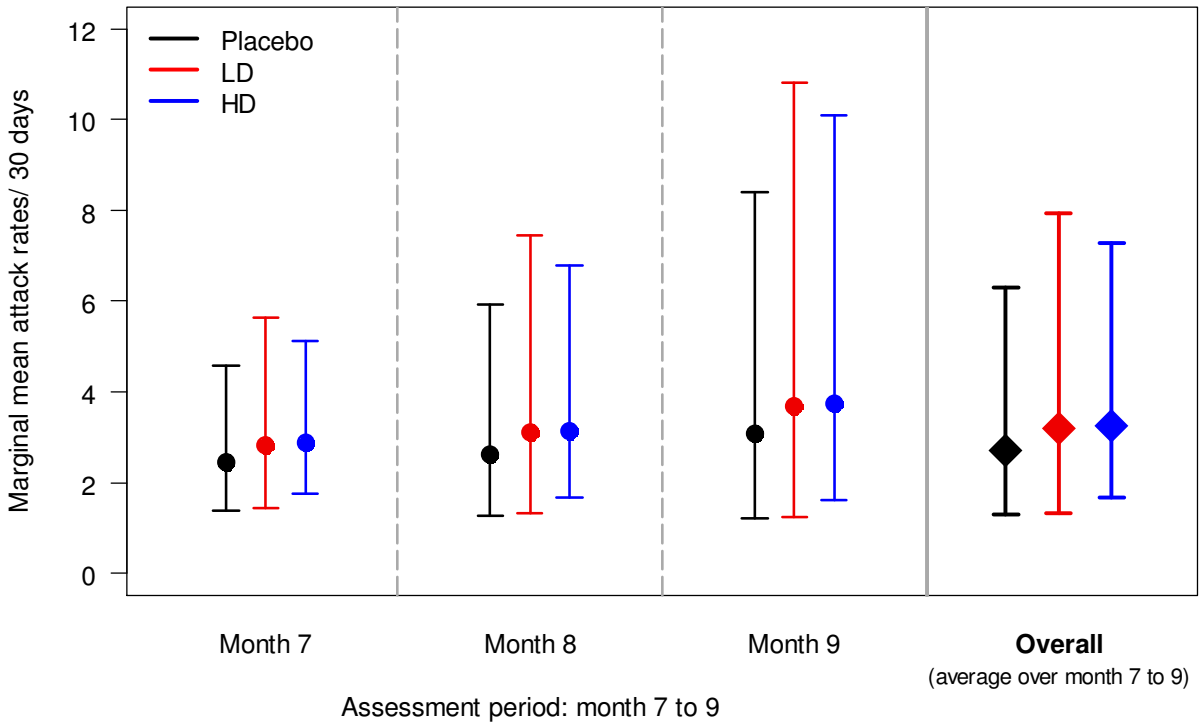

Figure S2. Marginal mean attack rates (with 95% CIs) per 30 days at month 7 to 9, and averaged over the 3-month long assessment period (filled diamond) for the FAS sample. Rates for each of the three treatments were estimated using a negative binomial mixed effects model based on all available data over the whole 9-month treatment period.

S3. Secondary Efficacy Outcomes

3.1 Definition of the Selected Ear

According to the inclusion criteria, a study participant suffers from audiometrically documented hearing loss in either the left or right ear, or both ears. Additionally, tinnitus or aural fullness in the treated ear has to be diagnosed prior to enrolment. The *selected ear* chosen for statistical analyses was defined as follows:

- For patients with audiometrically documented hearing loss either in the left or right ear, the selected ear is the ear with hearing loss.
- For patients with audiometrically documented hearing loss in both ears and documented tinnitus/aural fullness in either the left or right ear, the selected ear is the ear affected by tinnitus/aural fullness.
- For patients with audiometrically documented hearing loss in both ears and documented tinnitus/aural fullness in both ears, the selected ear will be chosen randomly.

This strategy avoids bias-away-from-null which would be the case if the 'most affected' ear had been defined, as in many MD trials.

3.2 Quality of Life (QoL): Dizziness and self-assessment questionnaires – Definition of Total Scores

3.2.1 VDADL score

To determine how well patients judged their functional compensation, they completed self-administered questionnaires designed for vestibular patients that included the vestibular disorders activities of daily living (VDADL) scale. The VDADL consists of 28 questions that assess subjects' comfort and ability to perform activities categorized as *functional (F), ambulatory (A),* and *instrumental (I)*, as well as a "total scale" that summarizes all three categories. In the original

definition of the VDADL, subjects score their responses to each question using integer numbers ranging from 1 ("`best`") to 10 ("`worst`").

According to Cohen & Kimball (2000) the parameter to summarize the three subscales and the total score is the median score. In this way, if the patient fails to answer a question, the VDADL score is not affected significantly by missing values. Unlike the mean, the median is not unduly influenced by extreme answers that do not agree with the remainder of the subject's assessment, and avoids the bias that would be introduced into a sum if a subject omits an answer or uses the non-applicable rating ("`NA`").

The *VDADL total score*, i.e. the median value of answers across all 28 questions, was used as secondary efficacy outcome.

3.2.2 DHI score

To assess the impact of impairment the patients were asked to fill out the 25-item DHI questionnaire. The original DHI total score (range: 0 to 100 points) consists of three subscales: *functional subscale (F), emotional subscale (E) and a physical subscale (P)*. The top score is 100 (maximum perceived disability), the bottom score is 0 (no perceived disability).

The subjective measure of the patient's perception of handicap due to the dizziness can be categorized as follows (Jacobson & Newman, 1990):

- 16–34 points (mild handicap)
- 36–52 points (moderate handicap)
- 54+ points (severe handicap).

For reach of the 25 items, a "`yes/always`" response is scored 4 points, a "`sometimes`" response 2 points, and a "`no`" response 0 points.

To deal with missing items, we used the derived *DHI mean total score* (DHI $\text{Total}_{\text{mean}}$) as outcome variable averaging for the number of answered questions:

$$\text{DHI Total}_{\text{mean}} = \left(1/\sum_i item_i \neq NA\right) \sum_{i=1}^{25} item_i$$

where `NA` denotes a missing answer. In `R` code this means: `mean(., na.rm = T)`.

3.2.3 MiniTF12 score

The full tinnitus questionnaire (TF) of Goebel and Hiller (1994) measures the impairment due to tinnitus with six partially correlating factors and is a standardized instrument for grading the severity of tinnitus.

Instead of using the full TF global score (in which 40 of the 52 items are needed for computation of the total score), the *MiniTF12 score* according to Hiller & Goebel (2004) as an abridged and more compact measure was analyzed to assess tinnitus-related psychological distress. The following selected 12 items reflect most central and characteristic aspects and are used to calculate the MiniTF12 score:

- [5] I am aware of the noises from the moment I get up to the moment I sleep.
- [16] Because of the noises I worry that there is something seriously wrong with my body.
- [17] If the noises continue my life will not be worth living.
- [24] I am more irritable with my family and friends because of the noises.
- [28] I worry that the noises might damage my physical health.
- [34] I find it harder to relax because of the noises.
- [35] My noises are often so bad that I cannot ignore them.
- [36] It takes me longer to get sleep because of the noises.
- [39] I am more liable to feel low because of the noises.
- [43] I often think about whether the noises will ever go away.
- [47] I am a victim of my noises.
- [48] The noises have affected my concentration.

Each item can be answered as either "`true`" (= 2 points), "`partly true`" (= 1 point) or "`not true`" (= 0 points). The crude MiniTF12 score is the sum of all points, ranging from 0 to 24.

As described in section 3.2.2, we used the derived *MiniTF mean total score* ($\text{MiniTF}_{\text{mean}}$) as an outcome variable, averaging for the number of answered questions defined above (item number #5, 16, 17, 24, 28, 34, 35, 36, 39, 43, 47, 48) and ignoring the missing values

$$\text{MiniTF}_{\text{mean}} = \left({}^{1}/_{\sum_i item_i \neq NA}\right) \sum_{i \in \{5,16,17,24,28,34,35,36,39,43,47,48\}} item_i$$

where *NA* denotes a missing answer. In R code this means: `mean(., na.rm = T)`.

3.3 Acoustic Evoked Potentials (AEP)

According to the trial protocol, "objective hearing loss" – determined by acoustic evoked potentials (AEP) – was a further secondary efficacy outcome. However, during the blind data review we realized that it was not possible to analyze these data due to extremely poor data quality and a huge amount of missing data. The main reason for the latter was that, too often, this examination was not performed at all. From a clinical point of view, AEPs do not capture Menière's-specific cochlear dysfunctions, and so would not have been a meaningful secondary outcome anyway.

3.4 Secondary efficacy endpoints: Complete case ANCOVA for absolute change

The following table shows the results of the complete case analyses for the secondary efficacy outcomes (QoL and observer-reported outcomes). The ANCOVA model for absolute change scores adjusting for baseline value contains main effects for treatment group and baseline values.

Following the concept of Fox (2003), the effects were calculated absorbing the lower-order terms marginal to the term in question (i.e. treatment group), and therefore averaging over the baseline term of a particular ANCOVA model.

Supplementary Table. Mean absolute change (95% CI) for secondary efficacy outcomes together with the number of patients with measurements at baseline and month 9 visit (FAS sample).

Absolute change (Month 9 – BL)	**PL (N = 72)**	**LD (N = 70)**	**HD (N = 72)**	***P* value***
Quality of life scores				
MiniTF, *mean* total score	-0.121 (-0.223 to -0.019)	-0.113 (-0.212 to -0.014)	-0.140 (-0.240 to -0.039)	0.929
N	54	58	56	
VDADL, total score	-0.202 (-0.405 to 0.000)	-0.261 (-0.461 to -0.060)	-0.360 (-0.560 to -0.159)	0.547
N	57	58	58	
DHI, *mean* total score	-0.497 (-0.689 to -0.305)	-0.364 (-0.554 to -0.173)	-0.515 (-0.705 to -0.325)	0.482
N	56	57	57	
***Tinnitus intensity* [dB]**	-0.558 (-6.024 to 4.9078)	7.066 (0.533 to 13.598)	-1.823 (-7.957 to 4.311)	0.107
N	35	24	28	
***Peak slow-phase velocity* [°/sec]**				
cool water irrigation (30°C)	-0.126 (-1.605 to 1.353)	-0.892 (-2.401 to 0.616)	0.489 (-1.020 to 1.999)	0.442
N	52	50	50	
warm water irrigation (44°C)	-0.107 (-1.824 to 1.611)	-1.676 (-3.443 to 0.090)	-1.044 (-2.811 to 0.722)	0.449
N	54	51	51	
***Pure-tone audiometry (bone conduction): hearing loss* [dB]**				
250 Hz	-5.533 (-9.010 to -2.057)	-1.986 (-5.195 to 1.224)	-2.883 (-6.119 to 0.352)	0.316
N	34	40	39	
500 Hz	-4.372 (-8.386 to -0.358)	0.288 (-3.636 to 4.212)	-3.268 (-7.099 to 0.563)	0.231
N	44	46	48	
1000 Hz	-5.441 (-9.206 to -1.677)	-0.600 (-4.287 to 3.088)	-2.956 (-6.680 to 0.769)	0.196
N	47	49	48	
2000 Hz	-1.534 (-4.937 to 1.869)	0.612 (-2.575 to 3.798)	-1.840 (-5.098 to 1.418)	0.513
N	45	51	49	

* Complete case ANCOVA for absolute change, with factor for treatment group, and baseline value of the dependent variable used as a covariate. P-values resulting from global testing (F-test).
Absolute change means difference of 9-month value minus baseline value.

For VDADL Total Score: absolute change in median score was analyzed.
Tinnitus intensity [dB], caloric irrigation and hearing loss assessed for the "selected ear". QoL scores, tinnitus intensity, caloric irrigation and hearing loss: higher values at time point BL or 9-month visit indicate more severe impairment; a negative value in absolute change means that impairment improved over time.

References

Bauer P. Multiple Testing in Clinical Trials. *Stat Med* 1991;**10**(6):871-89.

Cohen HS, Kimball KT. Development of the vestibular disorders activities of daily living scale. *Arch Otolaryngol Head Neck Surg* 2000;**126**(7):881-7.

Fox, J. Effect displays in R for generalised linear models. *Journal of Statistical Software* 2003;**8**(15):1–27. http://www.jstatsoft.org/v08/i15/.

Goebel G, Hiller, W. Tinnitus-Fragebogen (TF): Standardinstrument zur Graduierung des Tinnitus-Schweregrades - Ergebnisse einer Multicenterstudie. *HNO* 1994;**42**:166-172.

Hiller W, Goebe G. Rapid assessment of tinnitus-related psychological distress using the Mini-TQ; *Int J Audiol* 2004;**43**(10):600–604.

Jacobson GP, Newman CW: The development of the Dizziness Handicap Inventory. *Arch Otolaryngol Head Neck Surg* 1990;**116**(4):424–427.

Marcus R, Peritz E, Gabriel KR. On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika* 1976;**63**(3): 655–660.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/, 2014.